Finding and Understanding Influential Sets in Regression

Rollin Brant

University of Minnesota
School of Statistics
Technical Report #466

February 1986

University of Minnesota
School of Statistics
Department of Applied Statistics
St. Paul, Minnesota 55108

Supported by University of Minnesota
Single-Quarter Leave
Fall 1985

Finding and Understanding Influential Sets in Regression

by

Rollin Brant
Dept. of Applied Statistics
University of Minnesota
St. Paul, Minnesota 55108

SUMMARY

This paper addresses the problem of influential sets in linear regression. Past investigations into this area have tended to emphasize the computational difficulties associated with the identification of influential sets. Important conceptual difficulties, however, must be addressed in advance of computation. In particular, identification methods should facilitate subsequent interpretations and not merely provide an uninformative catalog of such sets. Likely interpretations of influential sets and relevant strategies based on clustering concepts are discussed.

1. Introduction.

In fitting linear models to data it is not unusual for a relatively small group of cases to play a disproportionately large role in determining the overall fit. While such an occurrence is not always undesirable, the identification of such "influential sets" will generally be of interest to an investigator, who, depending on context, must then make some judgment on the advisability of relying so heavily on just a few observations. Unfortunately, neither the customary inspection of residuals, nor the use of robust regression techniques provide reliable means for identifying influential sets. To this end, alternative diagnostic procedures have been developed, with the major emphasis being on the identification of influential cases. Attempts to generalize these procedures to uncover influential sets have met obstacles of a largely computational nature. In addition, fundamental difficulties associated with the interpretation of influential sets have not been adequately addressed. Here we shall seek methods that mitigate both kinds of difficulty.

We begin, in Sections 2 and 3, by providing a brief review of formal characterizations of influential sets and some of the available strategies for identifying them. Section 4 considers the problem of interpreting influential sets. In Section 5 we introduce identification methods based on clustering which facilitate such interpretation, and in Section 6 consider measures which are additionally useful in this regard.

The framework for investigation is as follows. Based on a sample of n elements from some population, observations on a response variable, y_i, and p explanatory variables, given as vectors $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, are recorded in pairs (y_i, x_i), $i=1,2,\ldots,n$, which are referred to as cases. It is assumed that the vector of response observations $y:(n\times 1)$ is related to

X:$(n\times p)$ the matrix of explanatory observations, $X=(x_{ij})$, by $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_p(\mathbf{0}, \sigma^2 I)$ is a vector of unobservable errors and $\boldsymbol{\beta}$:$(p\times 1)$ is a vector of unknown coefficients.

Though $\boldsymbol{\beta}$ is ostensibly the main target of investigation, the entire specification is invariably tentative and itself subject to scrutiny. After initial examination of the data based on simple descriptive measures and/or graphs, evaluation of the model usually begins with fitting the least squares estimate for $\boldsymbol{\beta}$, $\mathbf{b}$:$(p\times 1)$. This estimate in turn determines fitted values, $\hat{\mathbf{y}}=X\mathbf{b}$, residuals, $\mathbf{e}=\mathbf{y}-\hat{\mathbf{y}}$, and an estimate of σ^2, $s^2=\mathbf{e}^t\mathbf{e}/(n-p)$, which provide the customary basis for model calibration and criticism. In addition, the "Hat" matrix, $H=(h_{ij})=X(X^tX)^{-1}X^t$, so called since $\hat{\mathbf{y}}=H\mathbf{y}$, is an important component in diagnostic procedures.

The basis for many influence diagnostics is case and/or set deletion. Fundamental to measurement of the influence of a single case, i, are the case-deleted parameter estimates, $\mathbf{b}_{(i)}$ and $s^2_{(i)}$, which are the estimates obtained when the i'th case is omitted. Relative to the consideration of a set of cases, say $I=\{i_1,i_2,\ldots,i_m\}$, we analogously define $\mathbf{b}_{(I)}$ and $s^2_{(I)}$, estimates based on the data omitting the cases in I. In addition we shall let $\mathbf{y}_I$ and X_I denote the observations corresponding to cases in I, and set $H_I= X_I(X^tX)^{-1}X_I^t$, the associated submatrix of H. Summations shall be assumed to run from 1 to n unless otherwise noted.

2. Characterizing Influence

An excellent discussion of influence in general is provided by Cook and Weisberg (1982). The following brief review focusses on approaches to defining influential sets. This concept admits many possible formalizations, depending on the particular aims of model fitting.

Nonetheless, a relatively small number of "general purpose" diagnostic proposals have been made, most of which are simple extensions of single case diagnostics.

The diagnostic use of components of H is naturally motivated. Owing to the idempotence of H,

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2, \qquad (2.1)$$

revealing that if h_{ii} is near 1, $\hat{y}_i$ is principally determined by y_i, which will thus tend to be influential in the overall fit (Huber, 1977). Thus h_{ii} has been termed the leverage for case i. Analagously, measures of set leverage derive from applying matrix norms to H_I, the multi-case analogue of h_{ii}, which we consider in Section 6.

It is informative to note that h_{ii} satisfies (assuming the inclusion of a constant term)

$$nh_{ii} = 1 + (x_i - \bar{x}) S_X^{-1} (x_i - \bar{x})^t, \qquad (2.2)$$

where $\bar{x} = n^{-1} \sum_{i=1,n} x_i$ and $S_X = n^{-1} \sum_{i=1,n} (x_i - \bar{x})^t (x_i - \bar{x})$. Thus leverage is a consequence of x_i's remoteness relative to $\bar{x}$ measured by the Mahalanobis distance. Additionally one has $\sum h_{ii} = p$, so that the average leverage is p/n. A lower bound, $h_{ii} \geq n^{-1}$ derives from (2.2), while $h_{ii} \leq 1$ follows from (2.1). Based on these considerations, cutoffs of the form $h_{ii} > c \times p/n$, have been proposed for distinguishing leverage points (Hoaglin and Welsch, 1978; Velleman and Welsch, 1981).

A direct description of the impact on the overall fit of omitting an observation is provided by Cook's distance,

$$D_i = (ps^2)^{-1} (\mathbf{b}_{(i)} - \mathbf{b})^t X^t X (\mathbf{b}_{(i)} - \mathbf{b})$$

(Cook, 1977). For sets, this gives rise to the generalized distance, D_I, which, letting $\hat{\mathbf{y}}_{(I)} = X\mathbf{b}_{(I)}$ and $\mathbf{e}_{(I)} = \mathbf{y}_I - X_I \mathbf{b}_{(I)}$, can be expressed variously as

$$(ps^2)\, D_I = (b_{(I)}-b)^t X^t X (b_{(I)}-b)$$
$$= (\hat{y}_{(I)}-\hat{y})^t(\hat{y}_{(I)}-\hat{y})$$
$$= e_{(I)}^t H_I\, e_{(I)}.$$

The latter representation reveals that D_I is of an omnibus nature, combining $e_{(I)}$, which describes directly the discrepancy between the observations from set I and the fit derived from the remainder of the data, and H_I, which reflects the joint leverage of the cases in I. This dichotomy has important implications with regard to formal outlier tests. For simplicity's sake, consider the case that I={i}, a singleton, whence

$$D_i=(ps^2)^{-1}h_{ii}e_{(i)}^2.$$

In particular, consider the mean shift outlier model for the case i, which specifies as an alternative hypothesis, H_a: $E(y_i) = x_i\beta+\delta$, i.e. that $E(y_i)$ deviates from the nominal specification by some quantity δ. The customary test for this alternative is based on

$$t_i^2 = \{(s_{(i)}^2)^{-1}e_{(i)}^2\,(1-h_{ii})\} \sim F(1,n-p-1).$$

Significantly, the power of this test is smallest when h_{ii} is large. Thus D_i, and more generally, D_I, places emphasis on points that will be revealed as outliers in formal tests and on high potential sets whose validity is impossible to verify by conventional procedures. Such sets must therefore be evaluated in the light of other criteria, which will usually depend on the problem's specific context.

A measure of primarily geometric motivation, proposed by Andrews and Pregibon (1978), takes the form

$$R_I = \left|(X^*_{(I)})^t X^*_{(I)}\right| \times \left|(X^*)^t X^*\right|^{-1},$$

where $X^* = (X \mid y)$ and $X^*_{(I)}$ denotes the corresponding form, deleting set I. R_I is closely related to Wilk's test for outliers in sampling from Multivariate normal populations, in that it measures the outlyingness of

$\{(x_i, y_i), i \epsilon I\}$ in p+1 space. One apparent disadvantage is that it is invariant with respect to the designation of response among the p+1 variables, and thus does not reflect the response-explanatory dichotomy. The authors do however consider the determination of significance levels appropriate to the usual regression situation, where no distributional assumptions are made regarding **x**.

In Draper and John (1981) R_I is decomposed as

$$R_I = [1-\{(n-p)Q_I/s^2\}]\ |I-H_I|,$$

where $Q_I = \mathbf{e}_{(I)}^t(I-H_I)\mathbf{e}_{(I)}$ and I is the m×m identity matrix. The application of R_I, D_I, and the separate use of Q_I and $|I-H_I|$ are compared. Draper and John recommend the routine use of of D_I, Q_I and $|I-H_I|$.

Measures similar in form to D_I, have also been proposed, independently, by Belsley, Kuh, and Welsch (1980). In particular Welsch (1982) has recommended using

$$(n-m)\mathbf{e}_{(I)}^t X_I (X_{(I)}^t X_{(I)}) X_I\, \mathbf{e}_{(I)}/(ms_{(I)})^2.$$

Additionally an approach motivated by reference to the predictive use of the linear model, recommended by Johnson and Geisser (1983), can be straightforwardly generalized to multiple cases.

3. Identifying Influential Sets

3.1 Motivations

As noted previously, the motivation for considering influence measures is the insensitivity of the more familiar diagnostics to influential cases. In past investigations into the more general problem of influential sets, the chief concern has been the masking effect, which occurs when the real influence of particular cases is not discernible in

single case statistics due to the intervention of "masking" observations. The phenomenon is illustrated in Figure 1 by two basic configurations, A and B. For configurations such as these, the effect of deletion of any of the points separately in terms of the single cases measures is not indicative of the joint influence that the points exert. These effects are only discernible on deletion of the entire subset, indicating that subset deletion diagnostics are required to reliably detect all cases that are likely to be of interest.

The above phenomenon is illustrated in a data set arising from an investigation into the factors determining the selling price of houses, (Narula and Wellington, 1977, see also Weisberg, 1985). The variables considered were:

Y = sale price in dollars
X_1 = current taxes in dollars
X_2 = number of bathrooms
X_3 = lot size in square feet
X_4 = living space in square feet
X_5 = number of garage spaces
X_6 = number of rooms
X_7 = number of bedrooms
X_8 = age of house in years
X_9 = number of fireplaces

An initial fit yields the results given in Table 1. An index plot of the externally studentized residuals, $t_i=e_i/\{s_{(i)}(1-h_{ii})^{1/2}\}$, given in Figure 2 reveals no significant outliers. Plots of the single case statisics h_{ii}

and D_i in Figures 3, however, draw special attention to case 27, whose deletion results in the fit given in Table 2. Not apparent in any of the single case measures is the joint impact of deleting points 9 and 10. The tabulation of D_I for all pairs (see Table 3) reveals this as the most influential pair. The fit with the pair deleted is given in Table 4. Closer investigation reveals that these cases describe the largest and most expensive houses in the sample, and conceivably represent an untenable extrapolation of the model's assumption of approximate linearity. This can be formally verified by the fit of a single additional indicator variable corresponding to the pair, which yields an observed level of significance (two-sided) of p=.00001, confirming that the two taken together are not well described by the model considered.

The above illustrates that the identification of related influential cases facilitates the subsequent diagnosis of particular weaknesses in the model. In particular, the power of tests for localized inadequacy, typically low when based on single cases, increases when applied to relevant subsets of a larger size. As well, when cases are identified in groups it is much easier to perceive those sorts of patterns which suggest particular model improvements. Consequently, we see that the consideration of influential sets has two key motivations. The first is in overcoming the masking problem. The second aim is to provide some grouping of influential (singly or otherwise) points into groups as an aid to the criticism and augmentation of the model.

3.2 Methodology

The most direct approach to identification is to screen all possible sets according to a chosen measure and cut-off value. This is generally

infeasible due to relatively high expense of the individual computations relative to each set coupled with the large number of sets. To decrease the expense more selective screening can be done, say by restricting to sets of size $m \leq m_{max}$. In most cases, however, the evaluation a large number of sets may still be necessary. Thus one desires efficient, or at least, computationally feasible, algorithms for searching out the subsets with large values of particular measure of interest. Approaches to the computational problem have been considered by Andrews and Pregibon (1978), Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1980), Welsch (1982), and Gray and Ling (1984).

Andrews and Pregibon propose to mitigate the formidable screening problem by first applying the influence measures R_I to single cases only. They then consider only sets of high scoring cases, restricting further to sets of size $m \leq m_{max}$, where m_{max} is some reasonable bound chosen by the investigator. This approach alleviates, but does not avoid entirely, difficulties arising out of masking.

Screening all subsets of size $m \leq m_{max}$ can be facilitated by tree-search algorithms akin to those given by Furnival and Wilson (1974) in the context of variable subset selection. Such an approach to the calculation of set diagnostics is outlined in Belsley, Kuh and Welsch (1980). Unfortunately, the bounds on residual sums of squares employed by Furnival and Wilson to provide further shortcuts in variable selection, do not generalize to influence measures. Thus affordable "leaps and bounds" algorithms for identification do not seem achievable.

Cook and Weisberg (1980) consider the problem of determining all subsets of a fixed size m, whose D_I exceeds a given cutoff. They achieve an initial reduction in the number of sets to be considered by arguing that

finding an influential set which includes a proper subset provides little additional information. Thus, for instance, individually influential cases can be excluded from the candidate list for informative influential sets. In addition they give a number of bounds which substantially reduce the number of subsets for which D_I must be explicitly calculated. However as their sample computation attests, when the data are numerous, even these clever tricks serves to facilitate only consideration of subsets of size m=2 or 3.

A heuristic approach to the computationally feasible identification of influential subsets is given in Gray and Ling (1984), who base their approach on the augmented hat matrix,

$$H^* = X^*\{(X^*)^tX^*\}^{-1}(X^*)^t,$$

where $X^*=(X|y)$. The n×n matrix H^* is used as a similarity matrix in a clustering algorithm (specifically, k-clustering, see Ling, 1972) as a means of identifying subsets of potentially high influence. Additionally, the authors consider similar methods based on $-H^*$ and $M=(|h_{ij}|)$, which used in connection with the H^*-based method seem, according to experience with examples, to provide a fairly reliable method of screening subsets for potential influence according to a number of methods. Though empirically useful, the proposed methods lack a clear theoretical foundation, as pointed out in the discussion (Weisberg, 1984).

The main aim in the work described above has been to overcome the masking phenomenon. Due to inherent computational difficulties, we see that no truly practical methodology has emerged. More importantly, however, the methods considered above do little to serve our second, and perhaps fundamental, goal, which is the identification of substantively meaningful groups of observations. To seek methods which will better

serve this goal, we must consider more carefully what it is we can hope to learn from the identification of an influential set.

4. Interpreting Influential Subsets.

4.1 Influential Cases

However one chooses to define influence, some ambiguity arises with regard to the practical consequences of influential cases. The identification of a potentially anomolous case does not in and of itself indicate specific remedial action. Rather, it indicates the need for additional diagnostic measures aimed at investigating the following:

Case reliability. Have the associated observations been perturbed by (potentially correctable) gross errors, e.g. mistakes in recording the data? Is the case peculiar in some identifiable, if hitherto unsuspected, manner which warrants setting it aside for special treatment?

Model adequacy. Is the model defective in a systematic way, e.g. in missing terms in $X\beta$ or in the specification of the distributions of the ϵ? Does the case fall in a region of the predictor space where the assumption of approximate linearity is tenuous?

It is helpful if the above issues can be confronted in some orderly manner. A natural first step to take in regards to an anomolous case is to consider the possibility of purely haphazard errors, such as mistakes in data entry. If such can be ruled out, i.e. if the measurements appear valid to the best of the investigator's knowledge, more detailed examination is warranted. Coming to terms with the relevant issues above can be

facilitated by use of a corespondingly multi-faceted measure. As previously noted, writing Cook's distance as

$$D_i=(ps^2)^{-1}e_{(i)}^2h_{ii},$$

reflects the fact that, by this measure, influential cases will either be outliers, as indicated by $|e_{(i)}|$, leverage cases, or some combination both. Of course no simple recipe for action can be based this taxonomy, but the investigation of an influential case will be aided by the separate consideration of the components $e_{(i)}$and h_{ii}.

Since leverage points correspond essentially to outliers in the predictors, the occurrence of such cases provokes reflection on the likely tenability of the model in extreme regions of the predictor space. Also, such cases may often be deviant in other respects not accounted for by the model. The assessment of these possibilities is hampered, however, since as previously noted, testing procedures will usually lack power. In general, formal methods will be of limited utility, and context-dependent considerations will be key.

An apparent outlier also leads to considering if the offending case is unusual in some identifiable respect. If examination of the case uncovers peculiar characteristics not described by the predictors, a format outlier test is relevant and special treatment, e.g. case deletion, a plausible remedy. On the other hand, lacking substantial justification, the mere statistical signicance of an outlier based on the conventional assumption of normality casts as much doubt on that assumption as on the validity of the case. In such an case, accommodation, rather than separate treatment, is arguably the proper course.

Robust methods have been suggested as one form of accommodation. A not unrelated alternative course is to abandon the usual

parameterization in terms of conditional expectations, and adopt a more stable parameterization, for instance, in terms of conditional medians or trimmed means. This is especially appropriate if the assumption of symmetry is questionable. Many robust estimators can be interpreted in this light. For instance the one sample M-estimate, T, defined by the estimating equation

$$\sum\psi(y_i-T)=0$$

can be sensibly viewed as estimating the parameter μ defined by

$$\int\psi(y-\mu)dF=0$$

where F is the underlying distribution of y, avoiding the somewhat artificial assumption of symmetry.

4.2 Influential sets

Regarding interpretation of influential sets in general, one first notes that the inherent ambiguity that accompanies the identification of influential cases is likely to be exacerbated in the case of subsets. A related difficulty is the "swamping" phenonemon, which arises when the apparent high influence of a set is in fact attributable solely to a proper subset, e.g. one particularly influential observation. The problem is illustrated in Table 4 by the fact that the vast majority of apparently influential pairs contain case 27. Similar calculations for m=3 produce 20 sets with D_I exceeding 4 and 241 such sets for m=4.

Due to swamping, an inherent difficulty associated with the screening approach to identification of influential sets is that any tabulation of sets complete enough to include all potentially interesting sets will inevitably include many non-interesting subsets. Cook and Weisberg's approach to mitigating the problem is to omit sets which

themselves contain influential subsets. This provides an overly stringent solution, as pointed out by Welsch (1982), since substantively meaningful sets which happen to include singly influential cases will be overlooked. Thus, even if the computational problem of screening can be solved, there still remains the need for efficient means of examining the large number of subsets that screening tends to produce. This aspect of the problem seems to have been inadequately considered, perhaps because the screening problem needs itself to be solved before the swamping issue arises. However, filtering out the "relevant" influential sets poses a substantial obstacle to the profitable application of methodologies aimed at uncovering influential sets.

One strategy which mitigates both of the above problems consists of simply re-arranging the order of attack. Rather than screening all subsets and then looking for the meaningful ones, one may start with an initial reduction to potentially relevant sets, to which influence measures may then be applied. By relevant sets, we will generally mean sets having some more or less simple structure lending itself to interpretation with regard to influence related issues. The nature of such structure is considered below.

Recall that one of the two aims in finding influential sets is to relate influential observation in meaningful groups, i.e groups which have some common explanatory characteristic. Such characteristics may arise out of the predictors already in the model, or may stem from "lurking" variables. While the latter possibility can never be discounted, it is not an eventuality that is easy to anticipate in any formal methodology. Our efforts are likely to be more profitable if expended in the directions in which we have prior suspicions. One class of relevant sets to consider

first are those which are similar in some respect in the predictor variables, i.e. clusters in the predictor space. This intuitively plausible move can be given further heuristic support by the following argument. Suppose that a subset, I, of observations are subject to a similar bias relative to the model, δ, i.e. $E(y_I)=X_I\beta+\delta\mathbf{1}$, where $\mathbf{1}$ is a column vector of m "1"'s. Then $E(e_{(I)})=\delta\mathbf{1}$ and $e_{(I)}{}^tH_Ie_{(I)}$, the key factor in D_I will approximately equal $\delta^2\mathbf{1}^tH_I\mathbf{1}$. If $\bar{x}_I=m^{-1}\mathbf{1}^tX_I$, the average predictor vector, by letting $Z_I=X_I-\mathbf{1}\bar{x}_I$ one has that

$$\mathbf{1}^tH_I\mathbf{1} = m^2\bar{x}_I(m\bar{x}_I{}^t\bar{x}_I + Z_I{}^tZ_I + X_{(I)}{}^tX_{(I)})^{-1}\bar{x}_I$$

is maximized when the entries of Z_I are 0, i.e. when the cases in I are replicates. This suggests that large D_I values will tend to occur in clusters of exact or near replicates.

The second major aim, uncovering masked influential sets deserves reconsideration, as well. One notes that masking can either arise out of the more or less coincidental juxtaposition of anomolous observations, or can reflect systematic deficiencies in the model. The first possibility will by its own nature be somewhat of a rare occurence, albeit unfortunate, but thereby, one with which we cannot be overly concerned. The presence of systematic defects is an issue which must be paid greater attention. The typical masking configurations given in Figure 1 can be re-appraised in this light. The presence of an influential cluster, as in Configuration A suggests localized inadequacy in the model formulation, perhaps a departure from linearity in the extremes of the predictor space.

The occurrence of separated sets of mutually masking cases, as in Configuration B, has the aspect of coincidence. Of course such configurations may correspond to underlying relationships, which may be deduced by a sufficiently perceptive and/or persistent investigator. In

either case, however chance seems to play a significant role in connection with this sort of configuration, in that one must either be unlucky for them to occur, and/or rather lucky to be able to make any sense from them. On the other hand, the occurrence of high influence clusters is more likely to be substantially significant and informative.

By the arguments above, both major motivations in considering influential subsets will be well served by first considering sets which are clusters, with possible emphasis on the predictors. Such an initial reduction is a reasonable, if not, foolproof tactic, particularly, in light of the problems which arise if no such initial reduction is made. Of course, other sorts of meaningful structure, generally arising out of a problems peculiar context, can also be addressed. Clustering, however, has been found to be a serviceable general purpose tool in elaborating structure in high-dimensional problems. In the next section we consider relevant clustering strategies.

5. Clustering for influence

5.1 Clustering

Clustering as a diagnoistic aid has been considered previously in the context of diagnostics for regression models (see Daniel & Wood, 1980), and extended to generalized linear models (Landwehr, Pregibon, & Shoemaker, 1984). In those places, hierarchical clustering algorithms were used to partition observations into near-replicates, for the purpose of computing lack-of-fit statistics.

The use of clustering in the present connection has already been considered by Gray and Ling (1984), whose methods seem to work in practice, without having a strong theoretical justification. The relevance

of their use of H^*, however, can be elucidated when one notes that its elements can be written (assuming the inclusion of a constant term) as,

$$n h^*_{ij} = 1 + (x^*_i - \overline{x^*})^t \, S_{X^*}^{-1} \, (x^*_j - \overline{x^*})$$

where $x^*_i = (x_i, y_i)^t$, $\overline{x^*} = n^{-1} \sum x^*_i$, and $S_{X^*} = n^{-1} \sum (x^*_i - \overline{x^*})^t (x^*_i - \overline{x^*})$. The h^*_{ij}'s relate the cases in terms of inner products with respect to their estimated covariance structure in the (p+1) observation space. As well, they can be related directly to the Mahalanobis distance,

$$d_M(x^*_i, x^*_j) = (x^*_i - x^*_j)^t S_{X^*}^{-1} (x^*_i - x^*_j),$$

by noting that $d_M(z_i, z_j) = h^*_{ii} + h^*_{jj} - 2h^*_{ij}$, which together with the above, implies that

$$n h^*_{ij} = .5 * \{ d_M(x^*_i, \overline{x^*}) + d_M(x^*_j, \overline{x^*}) - d_M(x^*_i, x^*_j) \} + 1.$$

The measure of similiarity afforded by the use of H^* thus takes into account the remoteness of cases considered separately, in addition to their proximity. Consequently, clustering on the matrix H^* tends to give rise to (potentially influential) outlying clusters relative to the Mahalanobis distance.

Though the above sheds some light on the apparent utility of the Gray and Ling method, the heuristic nature of the method offers little assistance in the subsequent interpretation of identified subsets. In particular their approach fails to take into account the response-explanatory distinction. The question then is what type of clustering is likely to best serve our ultimate aims.

5.2 Case distances.

Clustering methods in general have the following outline. Since the aim of clustering is to define sets of "similar" cases, the fundamental construct is the distance matrix, $\Delta = (\delta_{ij})$, which describes dissimilarities

between pairs of cases (i,j). Taking the simple view that cases correspond to (p+1)-variate observations (x_i, y_i), one can refer to the clustering literature to find any number of distance measures. According to our previous arguments, measures that reflect the response-explanatory dichotomy are required. One possibility is to base distances solely on the explanatory variables. An inherent benefit of this is that the sampling behaviour of subsequently derived measures, such as lack-of-fit statistics, will be easier to calibrate. This is the approach adopted in Landwehr, Pregibon and Shoemaker(1984), who use simple Euclidean distance based on the x's.

A number of pitfalls limit the utility of this strategy. Firstly, the lack of invariance under scale changes requires that the choice of scales be carefully thought out. Secondly, and perhaps more importantly, the relevance of the distance measure is all too easily corrupted by the inclusion of irrelevant or redundant variables. For this reason, Atkinson and McCullagh (1984) suggest that the distance be based on fitted values. The danger in this approach, of course, is that it is heavily reliant on the aptness of the fitted values, which depends strongly on the very model whose validity is in doubt.

A compromise strategy which mitigates, though cannot eliminate, the above difficulties, is given by Daniel and Wood in their motivating paper, which uses the distance, Δ_{DW}, with entries

$$\delta_{ij}{}^2 = \sum_{k=1\ldots p} \{b_k(x_{ik}-x_{jk})\}^2 .$$

This measure shares to some extent the defects and virtues of the above alternatives. The associated geometry is invariant under scale changes in the explanatory variables, though not under arbitary affine transformations of **x**. Irrelevant variables are downweighted in the

distance, with some attendant dependence on the reliablity of **b**. This is not altogether undesirable, for if a particular coefficient is spuriously large due to the influence of a cluster of cases, such clusters are likely to emerge using δ_{ij} as the distance measure. If however, a variables effect has been masked, due to such a cluster, the methods based on Δ_{DW} may tend to overlook this sort of structure. This can be mitigated by the incorporation of a robust, bounded influence form of **b** in δ_{ij}, which will be sensitive to this sort of structure, (but then by the same token, be insensitive to that mentioned just previously).

The above measures by no means comprise an exhaustive catalog. In particular applications, substantial considerations may indicate more appropriate distance measures. In formulating a diagnostic approach, however, realistic limits must be placed on our ambitions. As a good all purpose distance measure, Δ_{DW}, represents a plausible compromise between our various, and possibly conflicting, aims.

5.3 Clustering approaches

Settling, at least provisionally, on the use of Δ_{DW} as the distance measure, it remains to choose among the various proposed clustering strategies (see e.g. Everitt, 1980). Hierarchical methods have received most attention in the statistical literature, and are implemented in the commonly available packages, thus seeming a natural choice. One difficulty, however, is the more or less arbitrary choice that needs to be made between the competing approaches, which include complete-linkage, single-linkage, average distance, and k-clustering (Ling, 1972) methods. That so many clustering strategies have emerged is a consequence of that fact that no single hierarchial method captures the diverse scope of

cluster structures. The narrow focus such clustering algorithms results largely from the restriction to a partition-like structure inherent in the hierarchical approach.

An approach which is not so rigid, but still tractable can be based on the use of near neighborhoods, described as follows. For a case i choose $j_1, j_2, \ldots, j_n$, such that $\delta_{ij_1} \leq \delta_{ij_2} \leq \ldots \leq \delta_{ij_n}$, giving rise to the nested sets, $\{i\}, \{i,j_2\}, \ldots, \{i,j_1,\ldots,j_n\}$. These near neighborhoods provide a complete characterization of the topology induced by δ. More importantly, they are natural candidates for consideration as high influence subsets, and provide a flexible basis for investigation of such sets which avoids the difficulties of partitioning strategies. Since smaller neighborhoods are of most interest, a judicious choice to restrict to sets of size m or smaller will generally be made, and influence measures computed for such sets. The basic algorithm is simple, and the calculation of influence measures can be streamlined by taking advantage of the nested structure of sets.

The results of calculation are naturally displayed using generalizations of the index plots used in dealing with single cases. To mitigate the swamping effect, it is most informative to plot successive differences in measures for consecutive nested neighborhoods, rather than the raw measures themselves. For instance, with reference to D_I differences of the form $d_{ik}=D_K-D_I$, where $I=\{i,\ldots,j_{k-1}\}$ and $K=\{i,\ldots,j_k\}$ can be plotted and the results displayed in superimposed index plots. As an example, in Figure 4 such a plot describes D_I for Δ_{DW} based neighborhoods up to m=5 from the house price data. Referring to Table 5, which describes the relevant neighborhoods, the joint influence of cases 9 and 10 becomes apparent.

The real potential of the above method is in application to a larger data sets, where competing methods tend to be extremely cumbersome or lose sensitivity. Consider for example the gasoline vapor data referred to in Cook and Weisberg (1980) and described more fully in Weisberg (1985). The data consists of 125 cases describing the results of an experiment aimed at relating the quantity of vapors released on the filling of a gasoline tank (y), to initial tank temperature (x_1), gasoline temperature (x_2), initial vapour pressure (x_3), and vapour pressure of the gasoline (x_4). Single case measures give some indication of potential problems in Figures 5 and 6. Investigation of larger size subsets was considered by Cook and Weisberg, from a purely computational viewpoint, who illustrated the high computational cost of considering even small subsets based on an all possible subsets approach.

Figure 7 describes D_I for Δ_{pW} neighborhoods up to size 6. The plot indicates high influence neighborhoods in the vicinity of cases 76-77 and near cases 61-66. Closer examination of the neighborhoods tabulated in Table 6 reveals two apparently influential sets, one involving cases 61-65 and the other involving cases 58 and 73-77. The anomolous nature of these sets can be assessed further by inclusion of indicator variables for the two sets in question. The relevant t-tests (2-side) yield observed levels of significance smaller than .001, indicating apparent lack of fit. On closer examination the cases were found to be among a group of 17 which had unusually high values of X_1. Fitting a separate model to this distinct group yielded a result significantly different from the fit of the remaining observations, indicating a breakdown of the model in this isolated region of the predictor space.

6. Leverage and Outlier sets

6.1 Leverage sets

The above proposed strategy is effective in identifying potentially interesting influential sets. Further understanding of the precise implications of individual sets can be augmented by consideration of the leverage-outlier dichotomy. We begin by considering the measurement of leverage for a set.

From an algebraic point of view, at least, H_I is the natural analogue of h_{ii} when dealing with subsets. Since its matrix form is somewhat inconvenient, we are lead to consider its reduction to a scalar quantity. From the viewpoint of comparing subsets within a model, D_I is determined by the quadratic form, $e_{(I)}{}^t H_I\, e_{(I)}$, and whereas $e_{(I)} \sim N(0,\sigma^2(I-H_I)^{-1})$, a natural measure of leverage is

$$\gamma_I = \sup \{ e \in R^m \mid (e^t H_I\, e)/(e^t(I-H_I)e) \}.$$

It is easily seen that $\gamma_I = h_I/(1-h_I)$ (see Cook and Weisberg, 1982, p.141) where h_I is the largest eigenvalue of H_I, so that h_I represents a convenient extension of the single case leverage, h_{ii}, to subsets.

Just as h_{ii} can be considered on its own as an influence measure, one might consider wholesale screening for high leverage sets. The wholesale calculation of h_I for all subsets is computationally impractical, and as the following reveals, unnecessary. Suppose I is the disjoint union of sets J and K, of sizes m_J and m_K, respectively, with $m=m_J+m_K$. Then $h_I \geq \min(h_J,h_K)$ and, more importantly, $h_I \leq h_J+h_K$, implying that $h_I/m \leq \max(h_J/m_J, h_K/m_K)$. By the latter we see that the union of two low leverage subsets cannot give rise to a high leverage subset, and thus that masking is not a serious issue in looking for high leverage sets. Indeed by the simple extension of the above upper bound on h_I, we have $h_I \leq \sum_{i \in I} h_{ii}$

(Cook and Weisberg 1982, p. 146) and consequently, $h_I/m \leq \max(h_{ii}, i \in I)$, so that any high leverage set must contain leverage points. This lower bound shows that a high leverage case will tend to give rise to a large number of apparently influential subsets, due to "swamping". Thus, screening all subsets for apparently high leverage sets does not appear necessary, nor even desirable.

Some guidelines are required for assessing magnitudes of h_I. Firstly, some allowance for subset size is necessary. One notes that if I consists of m exact replicates then $h_I = m h_{ii}$, where h_{ii} is the common case leverage. Thus, subsets of differing sizes can reasonably be compared in terms of h_I/m, where m is subset size. Additionally, since for any partition of the data, $\Pi = \{I_1, I_2, \ldots, I_k\}$, the bound $\sum_{I \in \Pi} h_I \leq p$ holds, informal cutoffs of the form $h_I/m > c \times p/n$, generalizing the single case guidelines, seem reasonable. Additionally one notes that the Euclidean length norm,

$$\tau_I = \tau(H_I) = (\text{trace}(H_I^t H_I))^{1/2} = \left[\sum_{i,j \in I} h_{ij}^2 \right]^{1/2},$$

provides a useful and computationally convenient bound, $h_I \leq \tau_I$, if one wishes to avoid eigenvalue calculations.

It is convenient to apply h_I to nearest neighbor clusters and plot successive differences for nested neighborhoods in superimposed index plots. In Figure 8, set leverages, h_I, are described for the nearest neighbor systems of the house price data and the gas vapour data. In the house price plot, no unusally leveraged sets emerge, while in the vapour plot, the high leverage of the already noted influential sets is revealed.

6.2 Outlier subsets.

The remaining component of D_I, $\mathbf{e}_{(I)}$, carries the information regarding the response, and hence is more directly informative of discrepant behaviour with respect to the model. Assessment of the apparent magnitude of this discrepancy is based on $Q_I=\mathbf{e}_{(I)}^t(\mathbf{I}-H_I)\mathbf{e}_{(I)}$, owing to the fact that, under the model, $Var(\mathbf{e}_{(I)}) = \sigma^2(\mathbf{I}-H_I)^{-1}$. Our investigation in the previous section gives us that

$$D_I \leq (ps^2)^{-1} h_I Q_I/(1-h_I).$$

The use of Q_I in screening outlier subsets has been considered by Gentleman and Wilk (1975) and by Draper and John (1981). As previously indicated, Q_I can be used as the basis for formal tests for mean shift alternatives, $E(\mathbf{y}_{(I)}) = X_I\boldsymbol{\beta} + \boldsymbol{\delta}$. One has under the preliminary model that

$$F_I = Q_I\{ms^2_{(I)}\}^{-1} \sim F(m,n-p-m).$$

Large values of F_I provide evidence that I contains anomalous cases, specifically that one or more of the cases is not adequately described by the model.

The above statistic is best used as a means of flagging observations for subsequent consideration on more substantial grounds, rather than in any "automatic" outlier rejecting procedure. As noted previously, the statistical significance any test of H_0: $\boldsymbol{\delta}=\mathbf{0}$ calibrated under the Gaussian assumption, is subject to a number of interpretations: to the skeptical, a significant result offers as much evidence against the Gaussian assumption as it does against H_0.

Following the general strategy adopted here, Q_I, can be calculated for relevant clsters, and plotted in the superimposed index plots. Figure 9 gives plots for the examples considered previously. In the house price example, cases 9 and 10 stand out as an outlier pair, whereas in the

gasoline vapour data, no such characterization seems to apply to the detected influential sets. In the latter case, the high leverage of the identified set is apparently the key factor.

7. Conclusion.

The main contention here is that the problem of dealing with influential subsets is substantially more than the computational problem of identifying them, in that useful methods must facilitate the eventual interpretation of such sets. The suggested approach is founded on first considering what types of influential subsets are amenable to interpretation. While no single and canonical approach presents itself in this connection, a number of plausible and computationally tractable methods can be considered. In particular, the use of a simple measures of distance between cases, in combination with either of hierarchical clustering or nearest neighborhoods methods, has been shown to be useful in identifying subsets which are worthy of further consideration.

As increased computing power becomes available, the computational problems which prevent wholesale screening may be mitigated. However, the identification of meaningful structure in a catalog of apparently high influence subsets itself requires a large scale effort. The development of expert systems may hold some promise in this regard, but before this can be accomplished, formalization of the notion of meaningful sets in some relevant and context dependent fashion will clearly be required.

Many approaches beyond those considered here are of interest. Further advances may follow from projection pursuit type approaches, assuming that suitable figures of merit for influential cases can be defined, and incorporated in screening algorithms along with the more

fundamental measures. It is clear that a great variety of intriguing and challenging puzzles remain to be posed, let alone solved, along the way towards defining a truly reliable methodology for regression.

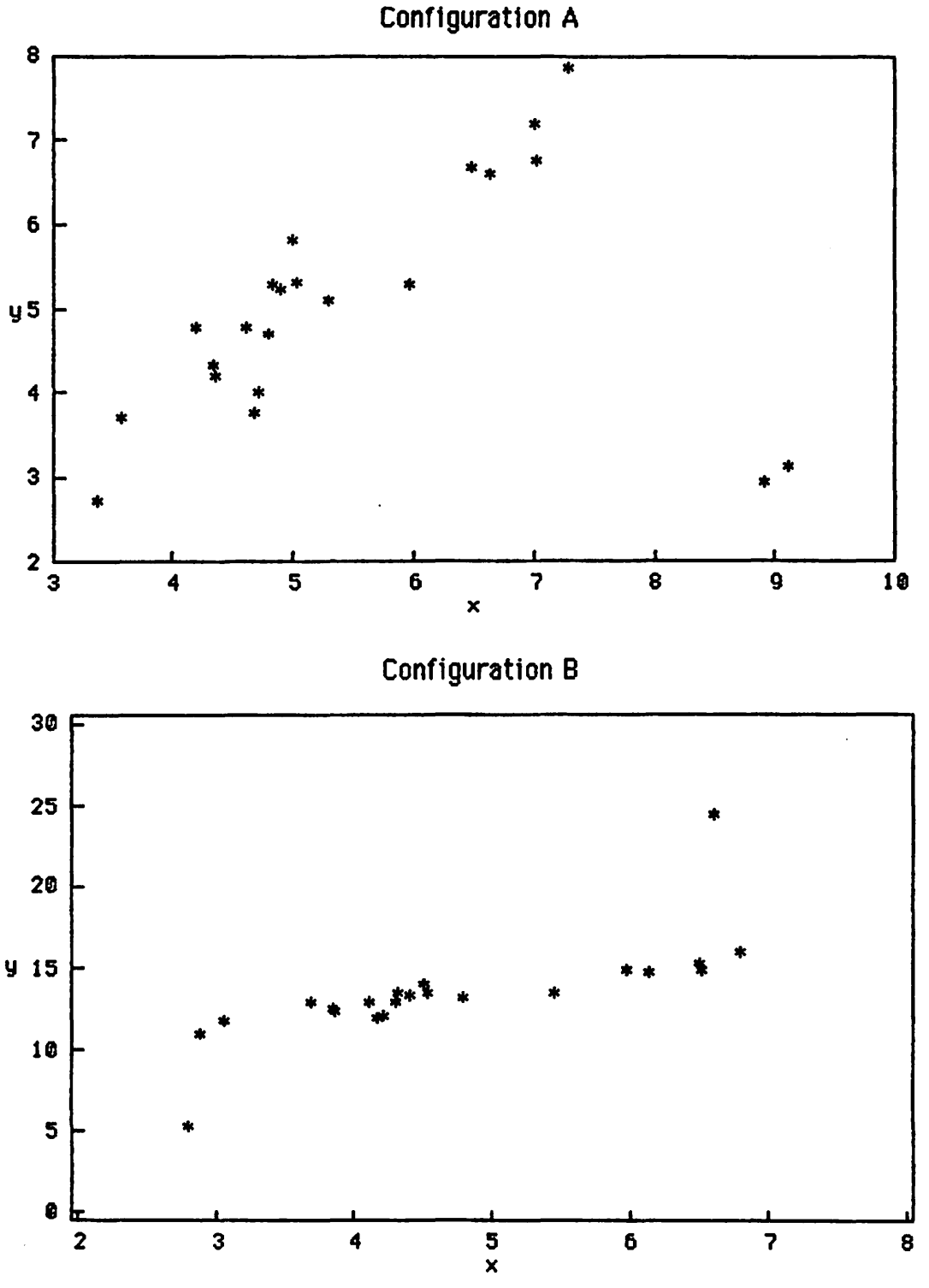

Figure 1. The two plots above portrary typical masking configurations.

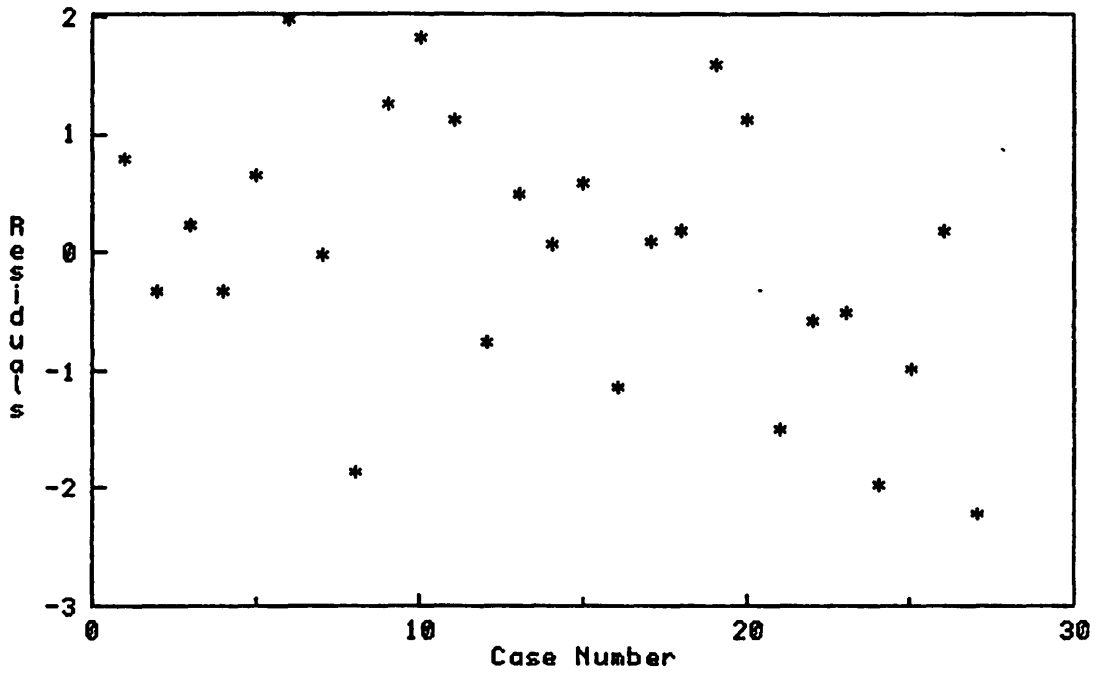

Figure 2. House price example - Index plot of externally studentized residuals.

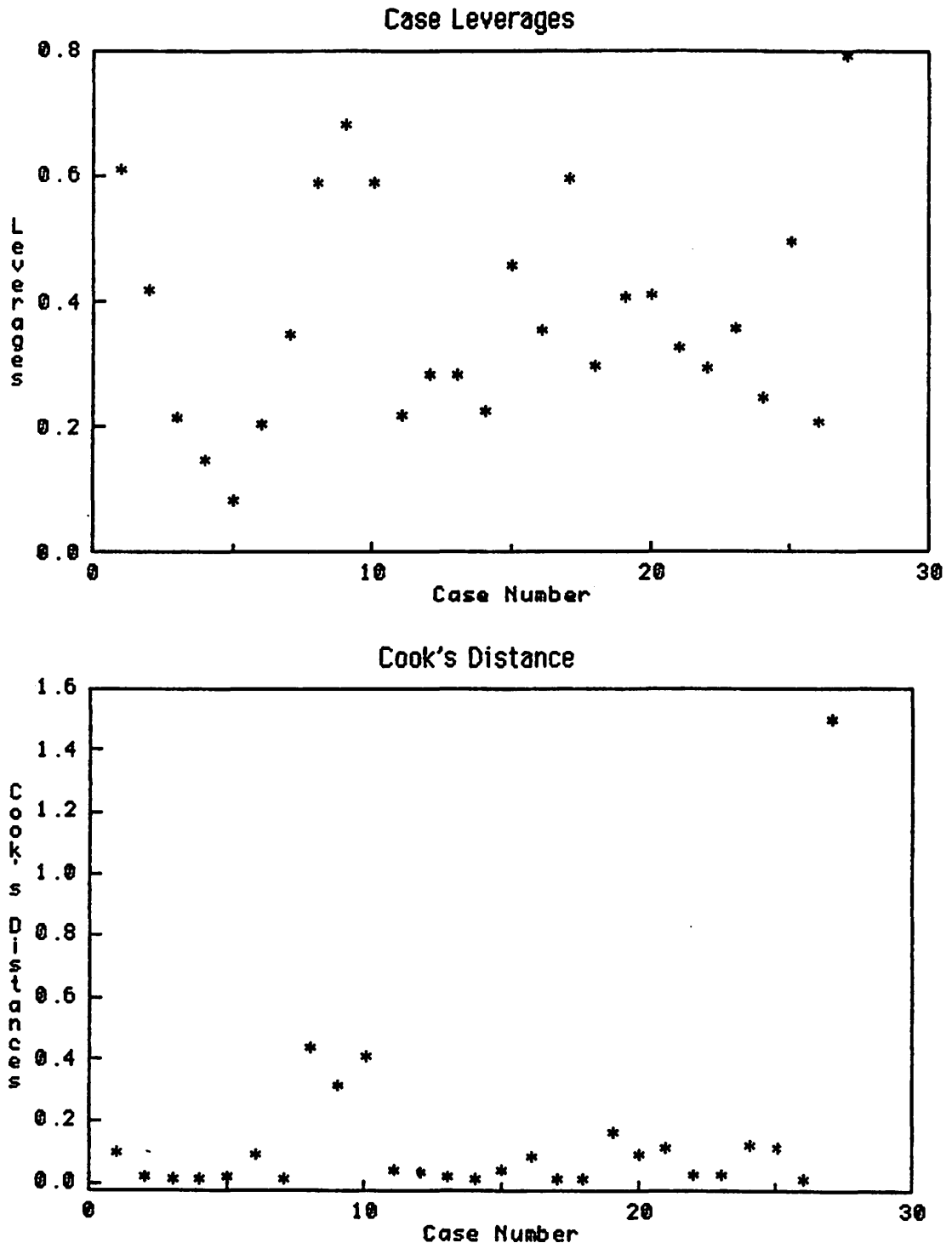

Figure 3. House price example - Case diagnostics.

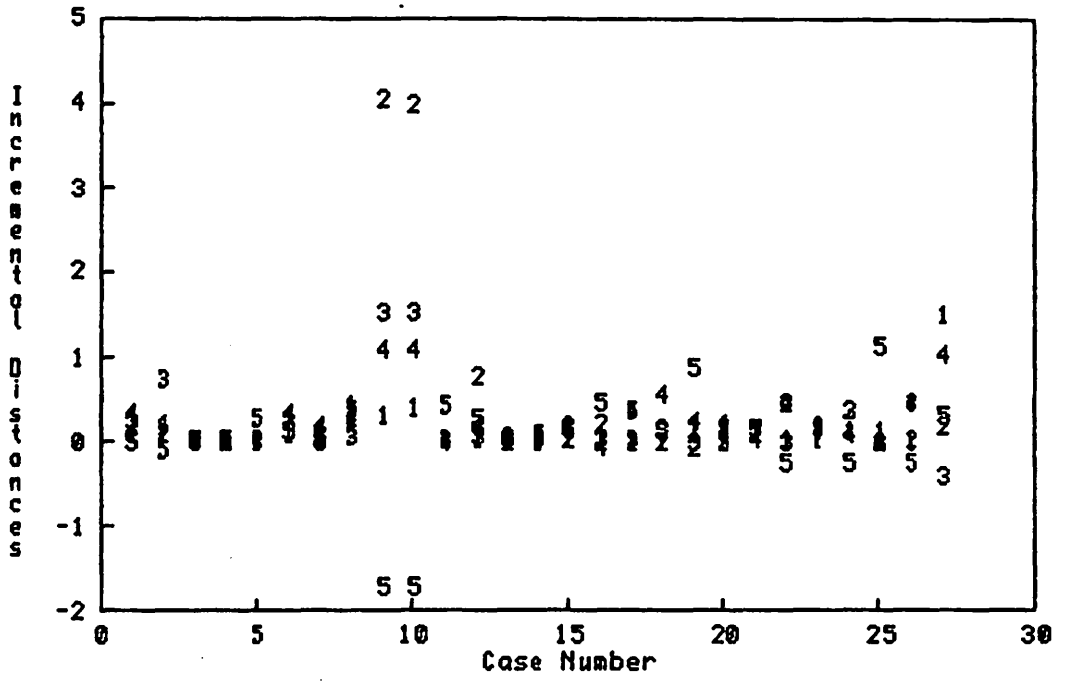

Figure 4. House price example - Superimposed index plot of successive differences, d_k, in Cook's distances for nested neighborhoods based on Δ_{DW}, k=1,...5.

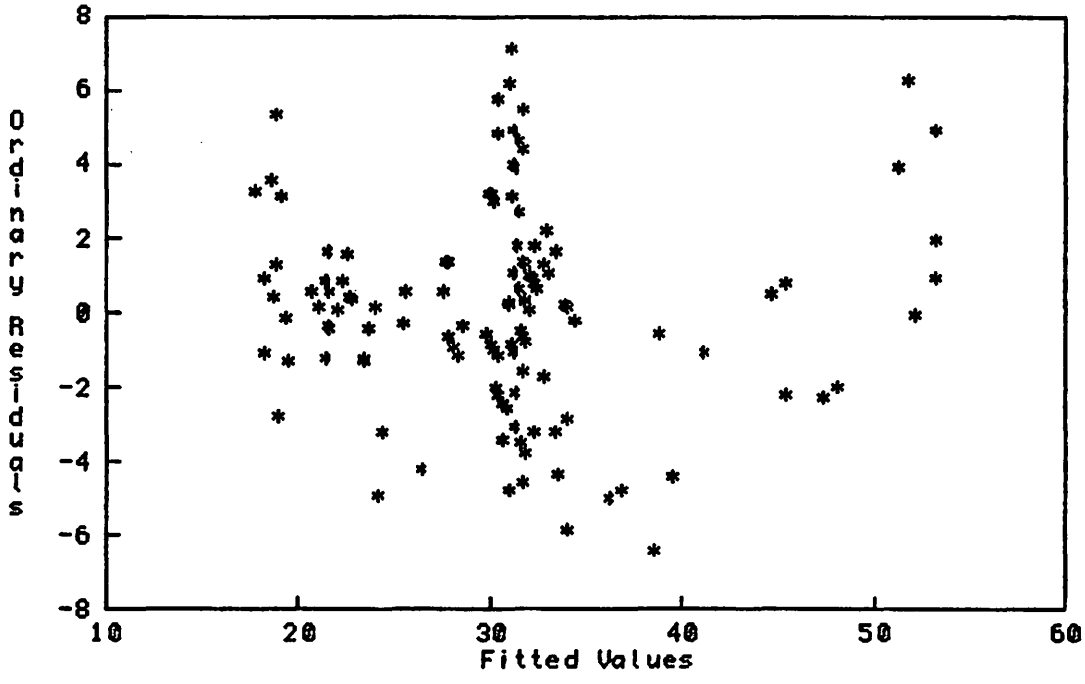

Figure 5. Gasoline vapor example - residual plot, ordinary residuals versus fitted values, from regression of y on x_1-x_4. Particular cases are identified by number.

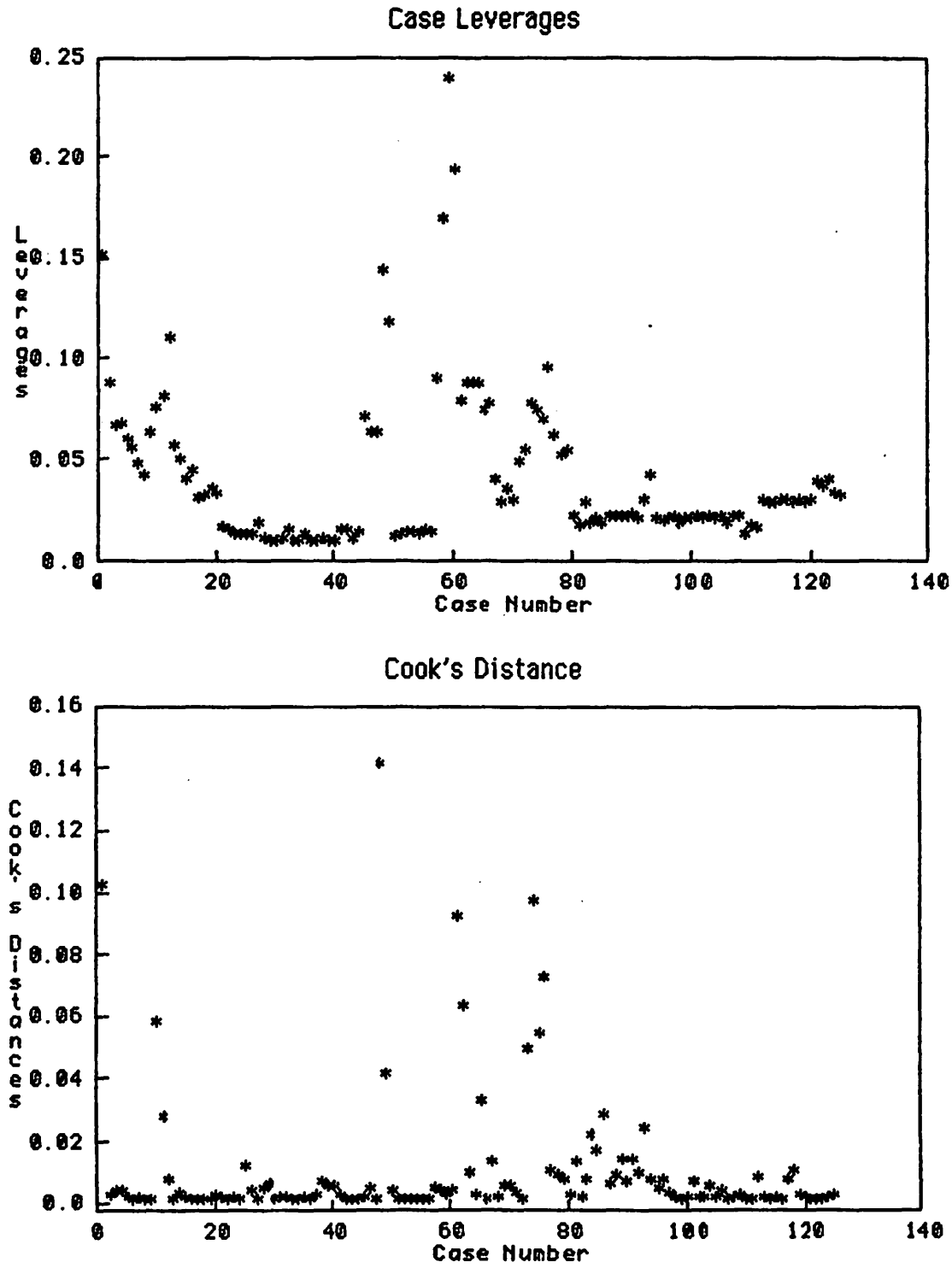

Figure 6. Gasoline vapor example - Case diagnostics

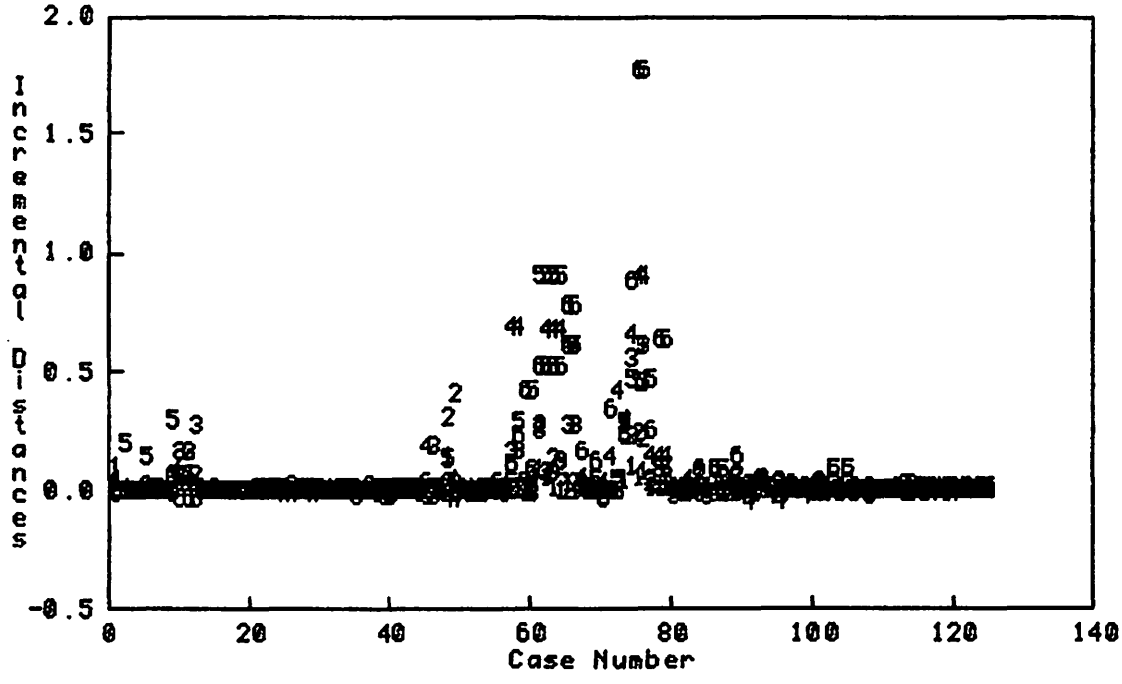

Figure 7. Gasoline vapor example - superimposed index plots of successive differences, d_{ik}, in generalized Cook's distance for nested neighborhoods based on Δ_{DW}, k=1,...6.

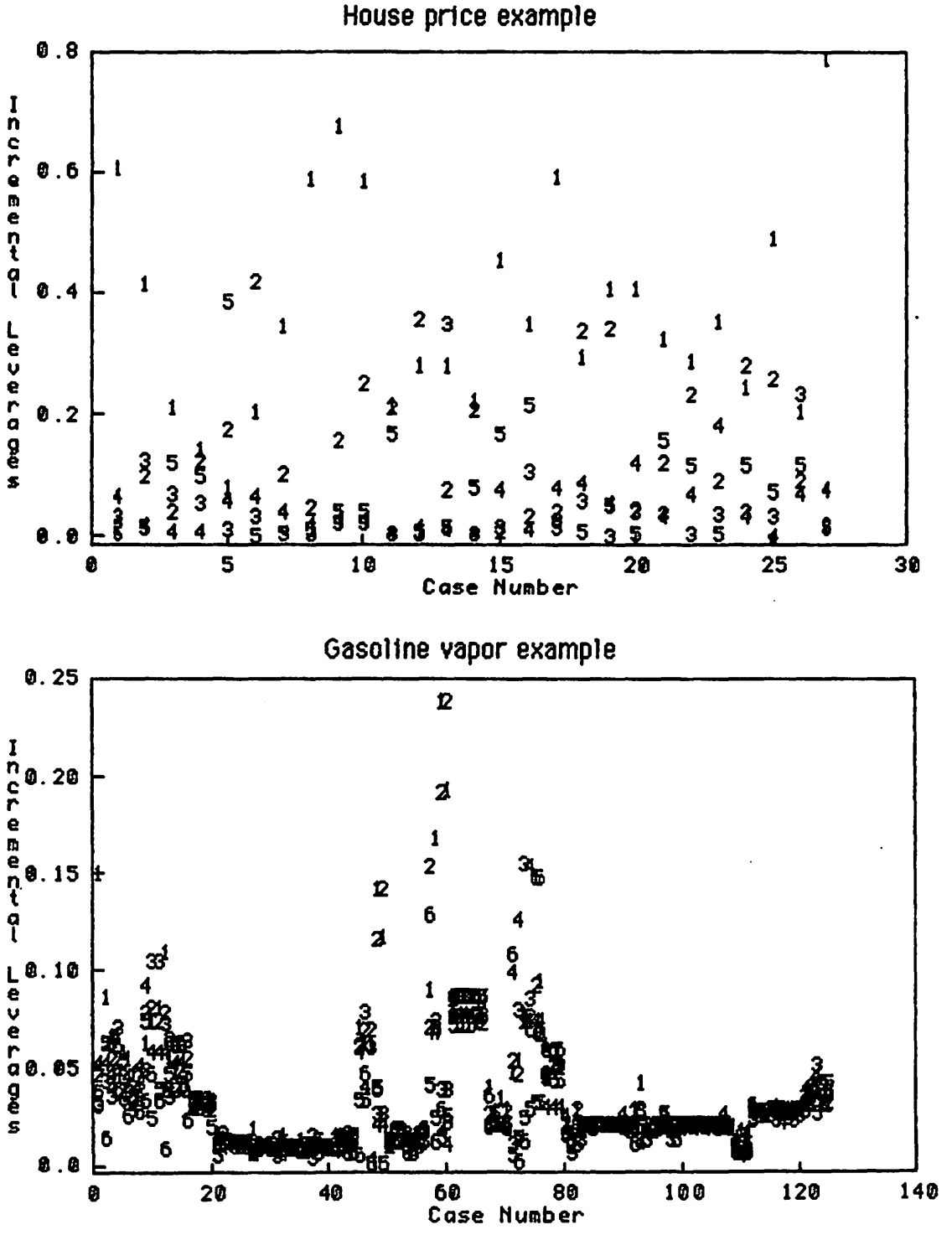

Figure 8. Set leverages (increments) for near neighborhoods based on Δ_{DW}.

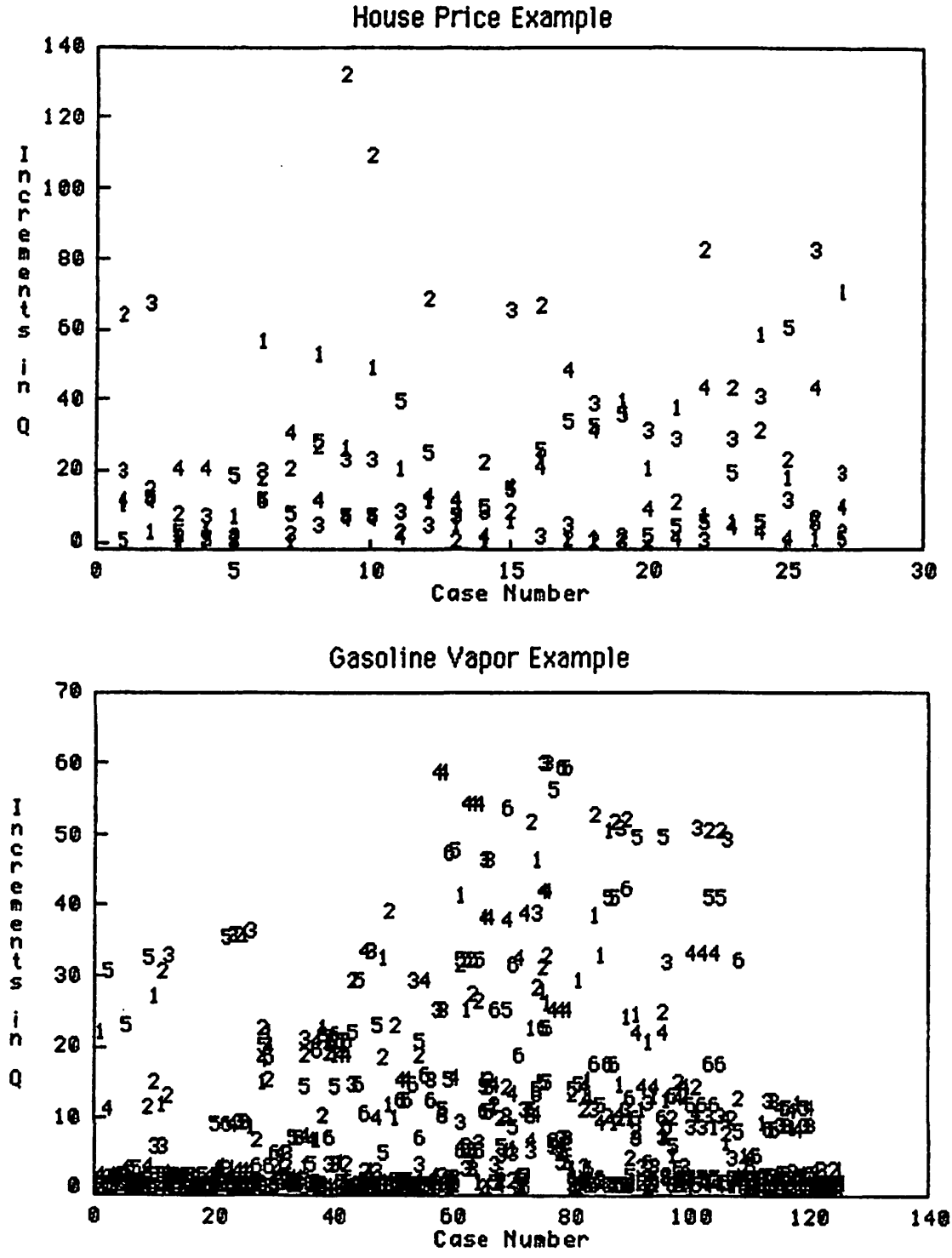

Figure 9. Superimposed index plots of successive increments in Q_I for nested neighborhoods based on Δ_{DW}.

Table 1. House price example - Initial fit.

	Coef	Std Err	t Value
Intercept	6.075697	7.239311	0.8392645
TAXES	1.235185	0.7899085	1.563706
BATHROOMS	7.314949	5.875446	1.245003
LOT SIZE	0.1902708	0.5620011	0.3385596
LIVING SPACE	13.47302	4.611445	2.921649
GARAGE SPACES	1.178733	1.887768	0.6244059
ROOMS	-0.7981690	2.419271	-0.3299211
BEDROOMS	-0.6265635	3.631443	-0.1725384
AGE	-0.0657891	0.0853271	-0.7710222
FIREPLACES	2.183505	2.415502	0.9039548

N = 27
Residual Standard Error = 4.16123
Multiple R-Square = 0.944691
F Value = 32.263 on 9, 17 df

Table 2. House price example - Fit after deleting case 27.

	Coef	Std Err	t Value
Intercept	12.47016	6.894323	1.808758
TAXES	3.339125	1.141679	2.924749
BATHROOMS	5.321845	5.197185	1.023985
LOT SIZE	-0.1949443	0.5177904	-0.3764926
LIVING SPACE	8.029766	4.661284	1.722651
GARAGE SPACES	1.019622	1.648135	0.6186522
ROOMS	-3.833121	2.485514	-1.542184
BEDROOMS	2.674934	3.474883	0.7697910
AGE	-0.0258243	0.07641286	-0.3379576
FIREPLACES	3.628789	2.197873	1.651045

N = 26
Residual Standard Error = 3.629835
Multiple R-Square = 0.957864
F Value = 42.93965 on 9, 17 df

Table 3. House price example - pairs of cases with D_I exceeding 1.

Pair #	Cases		D_I
1.	1	27	1.378443
2.	2	27	1.460718
3.	3	27	1.499184
4.	4	27	1.455401
5.	5	27	1.434071
6.	6	27	1.982067
7.	7	27	1.487367
8.	8	9	1.099254
9.	8	17	1.153594
10.	9	10	4.369832
11.	9	27	1.542622
12.	10	27	1.139819
13.	11	27	1.470820
14.	12	27	1.363460
15.	13	27	1.703831
16.	14	27	1.535012
17.	15	27	1.618515
18.	16	27	1.684534
19.	17	27	3.083667
20.	18	27	1.671008
21.	19	27	1.106483
22.	20	27	1.237749
23.	21	27	2.131597
24.	22	27	1.550358
25.	23	27	1.531716
26.	24	27	1.482888
27.	25	27	1.285607
28.	26	27	1.499126

Table 4. House price example - Fit after deleting cases 9 & 10.

	Coef	Std Err	t Value
Intercept	12.30116	5.327496	2.308995
TAXES	0.7901007	0.5622194	1.405324
BATHROOMS	8.141895	4.073024	1.998980
LOT SIZE	0.3892263	0.4154420	0.9368968
LIVING SPACE	4.394036	3.953067	1.111551
GARAGE SPACES	2.228946	1.307801	1.704346
ROOMS	1.459548	1.826292	0.7991868
BEDROOMS	-3.564105	2.856087	-1.247898
AGE	-0.05316637	0.06284146	-0.8460398
FIREPLACES	0.5415165	1.708715	0.3169144

N = 25
Residual Standard Error = 2.834428
Multiple R-Square = 0.842691
F Value = 10.11864 on 9, 17 df

Table 5. House price example - neighborhhood systems up to m=5 based on Δ_{DW}.

Case no.	Neighbors (in ascending order)			
1.	6	5	3	14
2.	12	8	4	11
3.	5	4	11	14
4.	3	5	11	7
5.	3	14	4	1
6.	1	5	3	15
7.	20	4	11	5
8.	12	2	4	25
9.	10	26	22	24
10.	9	26	22	24
11.	14	5	3	20
12.	8	2	23	25
13.	14	1	5	15
14.	11	5	3	13
15.	5	6	13	1
16.	24	26	21	22
17.	18	23	21	16
18.	17	21	25	19
19.	25	18	8	12
20.	7	11	14	4
21.	23	16	26	17
22.	24	26	16	21
23.	21	16	17	12
24.	22	16	26	21
25.	19	12	18	8
26.	22	24	16	21
27.	18	19	17	25

Table 6. Gasoline vapor example - neighborhood systems for cases 60-80, based on Δ_{DW}.

Case no.	Neighbors (in ascending order)				
60.	59	57	65	61	66
61.	62	63	64	65	66
62.	63	64	61	65	66
63.	62	64	61	65	66
64.	62	63	61	65	66
65.	66	61	62	63	64
66.	65	61	62	63	64
67.	92	80	94	99	93
68.	90	97	107	70	108
69.	88	101	84	89	86
70.	68	97	96	90	85
71.	72	5	48	13	49
72.	71	49	48	5	13
73.	74	58	57	77	78
74.	73	76	58	77	57
75.	76	74	73	77	58
76.	75	74	73	77	58
77.	78	79	73	74	57
78.	79	77	73	57	74
79.	78	77	73	57	74
80.	94	99	82	83	98

REFERENCES

Andrews, D.F. and Pregibon, D. (1978). "Finding the outliers that matter," *Journal of the Royal Statistical Society, Ser. B*, 40, 85-93

Atkinson, A.C. and McCullagh, P. (1984). Discussion to "Graphical Methods for Assessing Logistic Regression Models," by Landwehr, J.M., Pregibon, D. and Shoemaker, A.C., *Journal of the American Statistical Association*, 79, 72.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential and Sources of Collinearity*, New York: Wiley.

Cook, R.D. (1977). "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15-18.

Cook, R.D. and Weisberg, S. (1980). "Characterizations of an Empirical Influence Curve for Detecting Influential Cases in Regression," *Technometrics*, 22, 494-508.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, New York: Chapman and Hall.

Daniel, C.P. and Wood, F.S. (1980). *Fitting Equations to Data*. New York: Wiley.

Draper, N.R. and John, J.A. (1981). "Influential Observations and Outliers in Regression," *Technometrics*, 23, 21-26.

Everitt, B. (1980). *Cluster Analysis, 2nd ed.*, New York: Halsted Press.

Furnival, G. and Wilson, R. (1974). "Regression by Leaps and Bounds," *Technometrics*, 16, 499-511.

Gentleman, J.F. and Wilk, M.B. (1975). "Detecting Outliers II: Supplementing the direct analysis of residuals," *Biometrics*, 31, 387-410.

Gray, J.B. and Ling, R.F. (1984). "K-Clustering as a Detection Tool for Influential Subsets in Regression," *Technometrics*, 26, 305-318.

Hoaglin, D.C., and Welsch, R.E. (1978). "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17-22 and *Corrigenda* 32, 146.

Johnson, W. and Geisser, S. (1983). "A Predictive View of the Detection and Characterization of Influential Observations," *Journal of the American Statistical Association*, 78, 137-144.

Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984). "Graphical Methods for Assessing Logistic Regression Models," *Journal of the American Statistical Association*, 79, 61-71.

Ling, R.F. (1972). "On the Theory and Construction of K-Clusters," *Computer Journal*, 15, 326-332.

Narula, S.C. and Wellington, J.W. (1977). "Prediction, linear regression, and minimum sum of relative errors," *Technometrics*, 19, 185-190.

Velleman, P.F. and Welsch, R.E. (1981). "Efficient Computing of Regression Diagnostics," *The American Statistician* 35, 234-242.

Weisberg, S. (1984). Discussion of "K-Clustering as a Detection Tool for Influential Subsets in Regression," by Gray, J.B. and Ling, R.F., *Technometrics*, 26, 324-325.

Weisberg, S. (1985). *Applied Linear Regression 2nd ed.*, New York: Wiley.

Welsch, R.E. (1982). "Influence Functions and Regression Diagnostics," in *Modern Data Analysis* (Launer, R.L. and Siegel, A.F., eds.) New York: Academic Press, 149-170.